DOCUMENT RESUME

ED 413 360 TM 027 699

AUTHOR	Roberts, Lily
TITLE	Evaluating Teacher Professional Development: Local Assessment Moderation and the Challenge of Multisite Evaluation.
INSTITUTION	American Educational Research Association, Washington, DC.
SPONS AGENCY	National Science Foundation, Arlington, VA.; National Center for Education Statistics (ED), Washington, DC.
PUB DATE	1997-07-00
NOTE	37p.; Paper presented at the Annual Meeting of the National Evaluation Institute (Indianapolis, IN, July 9, 1997).
CONTRACT	RED-9255347; MDR9252906
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Attitude Change; *Educational Assessment; Educational Change; Intermediate Grades; Junior High Schools; Middle Schools; *Professional Development; Program Evaluation; *Science Education; *Student Evaluation; *Teachers; Teaching Methods
IDENTIFIERS	Moderation; Multiple Site Studies; Reform Efforts

ABSTRACT

The local assessment moderation process is described in terms of Science Education for Public Understanding Program (SEPUP) teachers' roles in the process of student evaluation and the subsequent use of that data for program evaluation. Teachers engaged in local assessment moderation function as a community of judgment, and this task engagement serves as a medium for teacher change. The principal challenge in evaluating the effect of local assessment moderation on teacher professional development was the multisite nature of the SEPUP "Issues, Evidence, and You" field test of a middle school science curriculum with embedded assessment. Teachers in SEPUP centers from Alaska to Washington, DC participated, so that interpretation of findings and presentation of results was complicated by differential organizational factors (leadership, institutional support, and teacher proximity and collaboration) and small sample size at the group level. Teachers from four Assessment Development Centers participated. The teacher change results were intriguing because they reflected a clear dissonance in teachers' minds about the rhetoric versus the reality of assessment reform. Teachers who used the assessment system were more likely to question its value in assessing learning, guiding instruction, and grading by the end of the year than those who were not required to use it. However, the latter group was more likely to use traditional assessment methods by the end of the year. In general the SEPUP Center most successful with local assessment moderation had higher means on measures of teacher change in assessment, collegial, and instructional practices than the less successful Centers. (Contains 4 figures, 3 tables, and 45 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 413 360

Evaluating Teacher Professional Development:

Local Assessment Moderation and the Challenge of Multisite Evaluation

Lily Roberts

University of California, Berkeley

July 1997

Paper presented at the annual meeting of the National Evaluation Institute. Indianapolis, July 9, 1997.

Panel Title:
Challenges in Science Education Assessment Reform:
Evaluating Teacher Professional Development and Student Change

TM027699

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Lily Roberts
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Abstract

The local assessment moderation process will be described in terms of the Science Education for Public Understanding Program (SEPUP) teachers' roles in the process of student evaluation and the subsequent use of that data for program evaluation. This process is integral to the concept of a *community of judgment* in terms of providing support for teacher change in assessment practices as well as ensuring quality control (in the technical sense, such as validity and reliability of scores). The principal challenge in evaluating the effect of local assessment moderation on teacher professional development was the multisite nature of the SEPUP IEY field test -- teachers in SEPUP Centers from Alaska to Washington, D.C. participated, so interpretation of findings and presentation of results was complicated by differential organizational factors (leadership; institutional support; and teacher proximity and collaboration) and small sample size at the group level.

The teacher change results were intriguing because they reflected a clear dissonance in teachers' minds about the rhetoric versus reality of assessment reform. Teachers who used the assessment system were more likely to question its value in assessing learning, guiding instruction and grading by the end of the year than those who were not required to use it. And yet this latter group was more likely to use traditional assessment methods by the end of the year (such as multiple choice questions) whereas the other group continued to use alternative forms of assessment. Cross-level exploratory analysis was used to examine differences in teacher change across SEPUP Centers. SEPUP Centers were ranked qualitatively by level of success with the local assessment moderation process, then within-Center mean scores were plotted by Center to explore relationships. In general, the most successful Center had much higher means on measures of teacher change in assessment, collegial and instructional practices than the less successful Centers.

Keywords: Multisite evaluation; science assessment; teacher professional development

Acknowledgments: This research was supported in part by a grant from the American Educational Research Association which receives funds for its "AERA Grants Program" from the National Science Foundation and the National Center for Education Statistics (U.S. Department of Education) under NSF Grant #RED-9255347 Further, this project has been supported by NSF grant No. MDR9252906, and also by other support through SEPUP at the Lawrence Hall of Science. Opinions reflect those of the author and do not necessarily reflect those of the granting agencies.

Evaluating Teacher Professional Development:

Local Assessment Moderation and the Challenge of Multisite Evaluation

Many educational reform efforts in the 1990's emphasize the link between teacher accountability and student performance (Cornett, 1995). Another significant feature of current efforts to improve education is an emphasis on assessment-driven reform (Linn, 1994; Shepard, 1995). Structuring a program to meet such expectations requires that complex approaches be included in a program's design. Examples include: (1) *project-based instruction* in which students investigate authentic problems in middle school science and teachers engage in "cycles of collaboration, enactment and reflection" to learn new strategies to change their classroom practices (Blumenfeld, et al., 1994; Krajcik et al., 1994); (2) the *Interactive Mathematics Project* (IMP) developed jointly by Lawrence Hall of Science and San Francisco State University which offers complete replacement courses for the traditional high school mathematics program and requires an extensive staff development program for teachers (Levine, 1994); and (3) the *SEPUP Assessment System* which is based on the integration of assessment throughout a new curriculum, and provides tools for teachers to assess student performance as well as evaluate their own instruction (Sloane, Wilson & Samson, 1996). Further, to meet the demands of state and federal agencies with goals of being first by the year 2000 in science and mathematics (e.g., National Educational Goals, see O'Sullivan, 1995, p. 19), project developers have designed complex treatments that endeavor to affect many components of the educational system simultaneously.

Challenges Presented by Complex Innovations

The complex, multilevel nature of approaches taken to improve mathematics or science education within a single project make it extremely difficult to disentangle program effects for evaluative purposes. For example, the use of multiple sites to enhance generalizability of program effects, results in projects crossing district and/or state lines, which only adds to the complexity. The evaluation of innovative programs is further complicated by context factors that exist at the locus of interaction between the teacher and students (i.e., the school) as well as at the nexus of the teacher and the organization (e.g., the district or the state). The multilevel aspects of teaching and learning can not be overlooked in any well designed evaluation.

Two common methodological challenges emerged in the evaluation of a middle school science curriculum and embedded assessment system developed by the Science Education for Public Understanding Program (SEPUP). The first methodological challenge is the need to make evaluation information meaningful and accessible to practitioners, including teachers, administrators, and curriculum and assessment developers (for a solution to this challenge, see Roberts, 1996, 1997).

The second challenge is to address the multilevel structure of complex innovations. This paper focuses on the latter methodological challenge, using the evaluation of SEPUP as a case study of the application of procedures used to provide substantive interpretations of evaluation information and to explore differences in program impact across multiple sites.

An Assessment System as a Medium for Teacher Change: Creating a Community of Judgment

The strategy of employing an integrated assessment system that addresses various elements of high quality professional development evolved from the concept that teachers need to form a *community of judgment* (Wilson, 1994) that can serve as a network much like the "professional-area movement organizations" described by Pennell and Firestone (1996). Wilson (1994) proposed the community of judgment as one approach to changing the "assessment culture" from one that relies primarily on standardized tests to a teacher-centered approach to educational accountability. Efforts to change the assessment culture are direct responses to calls for better assessments given the limitations of standardized tests (Bullough, 1988; Linn, 1987; Office of Technology Assessment, 1992; Shepard, 1989; Spring, 1988; Wiggins, 1989). To change the assessment culture, it seems clear that one must begin with the teacher who works directly with the students being measured. Shepard (1989) suggests that teacher enhancement is needed and that teachers should be directly involved in developing and scoring assessments. However, a comprehensive embedded assessment system goes far beyond the scoring task.

Wiggins (1989) suggested that any system of truly authentic assessment should meet several general criteria; the system should be: (a) criterion-referenced; (b) formative; (c) moderated; and (d) clear in the progression of educational development (e.g., score levels on a rubric would reflect the criteria). Wilson's (1994) *community of judgment* for educational accountability satisfies all of these criteria, while adding a fifth essential component -- a substantive *framework* that describes the achievement variables that are valued and thus worth assessing. Broad frameworks already exist in various subject areas, such as the *Benchmarks for Science Literacy* (AAAS, 1993), *Curriculum and Evaluation Standards for School Mathematics* (NCTM, 1989), or California's framework documents for mathematics, science, history/social science, visual and performing arts, foreign languages, physical education/health, and English/language arts. These broad frameworks serve as guidelines for local development and implementation of curriculum and assessment.

The substantive framework for assessment and accountability suggested by Wilson (1994) would require specific variables within a specific curriculum to be defined and measured, say, throughout a school year to chart students' progress. Tittle, Hecht, and Moore (1993) describe the use of an assessment framework for junior high school mathematics in which they studied students' (a)

metacognitive/self-regulatory skills and (b) affective, motivational and attributional beliefs. This assessment framework, however, is less robust than that proposed by Wilson because the framework is not specific to the substantive variables of the course, but rather a broad range of psychological constructs, such as "procedural skills, learning strategies, self-regulatory functions, and motivational orientations" (p. 14). Wilson's assessment framework specifies the constructs embedded in the course, such as the Designing and Conducting Investigations variable in the SEPUP middle school science course, *Issues, Evidence and You* -- thus assessment and instruction are conjoint rather than disparate elements of teaching. In their analysis of implementing the assessment framework, Tittle and her colleagues found that it was important to embed the assessment in "activity settings" and that teacher collaboration was critical to ensure that the assessments were meaningful as well as useful in classroom settings (p. 18). Tobin and Gallagher (1987) also found that "activity settings" (i.e., lecture format, whole class interaction, small groups, or individualization) in secondary science classes functioned to "mold the behavior of teachers and students" (p. 62). They found that high school science teachers in Australia and the United States geared their instruction to the "most able students" in the class (p. 74). Adjusting assessment strategies would hopefully counteract this non-equitable mode of teaching.

In addition to a substantive framework, the community of judgment (Wilson, 1994) model also contains (a) a set of *assessment modes* for each variable and scoring guidelines for rating responses; (b) *moderation* of the teachers' judgments that allows for rater improvement (i.e., teacher enhancement in assessment) and rater adjustment (i.e., consensus building to set standards for student performance); and (c) methods of *quality control* that ensure technical measurement standards (e.g., reliability and validity). Assessment modes are a variety of diverse indicators of student achievement, including tasks such as lab reports, performances (e.g., a town meeting) or projects. These modes are designed to promote higher order thinking, to engage students in real-world, meaningful activities, and to be as closely aligned as possible to the criteria of interest. Moderation is, in brief, a scorer-calibration process that can also serve as a mechanism for teacher professional development (Ingvarson, 1990; Linn, 1994; Roberts, Sloane, & Wilson, 1996). Quality control is needed to ensure comparability of teacher judgments across different settings and to authenticate student work (i.e., prevent cheating).

For teachers to assume a central role in this community of judgment for educational accountability, professional development is a must (Jett & Schafer, 1992; Linn, 1994; Newell, 1992); "teachers need better training and support as they face increasingly complex challenges of classroom assessment" (NCME, 1994). Particular to improving science education, good teaching materials, "vigorous and ongoing" professional development, and new "assessment tools" are basic components of successful reform (Lopez & Tuomi, 1995, p. 78). Richardson (1990) notes that "teachers themselves must be involved in making judgments about what change is worthwhile and significant" and "that practices and ways of thinking outside an

individual teacher's own experiences should be introduced into the dialogue" (p. 14). The community of judgment model can be used to engage teachers in making such judgments while benefiting from the interaction with colleagues.

Alternative science assessments will be implemented "only if teachers understand their use and the depth of the content they demand, are empowered to make instructional decisions, and are supported by school districts which encourage teacher change" (Harmon, 1995, p. 46). If teachers are engaged in a community of judgment for educational accountability, then it is far more likely that alternative assessments will be implemented and that they will be meaningful and useful in terms of informing teachers' instructional practices.

The greatest value of creating an assessment system that promotes teacher change is the potential impact it will have for students. Some benefits of revising and revitalizing assessment practices were reported by Roberts (1995) based on a statewide study of the Eisenhower Mathematics and Science Education Grant Programs implemented in California. Roberts noted that changing assessment practices enables teachers to focus on students' work and learning, so that:

- Teachers can diagnose student difficulties and provide appropriate support to scaffold learning and student understanding.
- Teachers can offer a greater variety of assessment activities to support the learning styles of a diversity of students.
- Teachers have evidence that supports their instructional decisions as well as enhances their communication with students, parents, administrators, and others.
- Teachers can fulfill a leadership role in the area of assessment (p. V-2).

Clearly, to meet the challenges of science education reform, the professionalization of teaching demands that teachers change not only their instructional practices, but also their assessment practices. Further, this change in teachers cannot and, in fact, will not occur in a vacuum; therefore providing for a professional community of judgment that serves as a medium for teacher change is essential.

SEPUP as a Case Study of a Complex Innovation

The Science Education for Public Understanding Program received funds from the National Science Foundation (NSF) to develop and field-test a year-long, middle school science course entitled *Issues, Evidence and You (IEY)*. In addition, funds were designated for the creation of an assessment system. The SEPUP Assessment Project was designed to apply new theories and methodologies in the field of assessment to the practice of teacher-managed, classroom-based assessment of student performance (Wilson & Adams, 1996).

One important feature of the assessment project is that the *assessment system is tied to the important learning developments* that are the goals of a specific curriculum (Sloane, Wilson & Samson, 1996). While compatible with other national, state, or district efforts to implement new forms of student assessment, the SEPUP Assessment System is specially designed for the specific curriculum that *IEY* teachers are using in their classrooms. A second important feature is that assessments are *fully embedded* in the instructional materials and are designed to be an integral part of the instructional process. A third important feature of the project is that the assessment of student performance, and the interpretation and use of information regarding student performance, is *managed by the classroom teacher*. A fourth important feature is that the assessment system is designed to provide teachers with *evidence* of their students' growth as well as the possibility of evaluating their own success.

The assessment system provides a set of tools for teachers to use to: (a) assess student performance on central concepts and skills in the curriculum; (b) set standards of student performance; (c) track student progress over the year on the central concepts; and (d) provide feedback (to themselves, students, administrators, parents, or other audiences) on student progress as well as on the effectiveness of the curriculum materials and classroom instruction. Initially, managing a new classroom-based system of embedded assessment demands much of the teacher. However, I believe that teachers must be recognized as the front-line professionals who will ultimately determine the usefulness or effectiveness of any attempt at educational reform. Empowering teachers (through the provision of tools, procedures, and support) to collect, interpret and present their own evidence regarding student performance is an important step in the continuing professionalization of teachers in the field of assessment.

SEPUP is an example of the kind of complex innovation that is currently being field tested in schools throughout the United States. The curriculum is designed to engage students in an "issue-oriented, hands-on" approach to thinking about scientific issues that are relevant to their daily lives (e.g., water, waste, energy, and environment) and their understanding is assessed in an ongoing manner. The students' role is changed as they conduct labs and other activities that are designed to help them understand that science is really a way of asking and answering questions and not just a collection of established facts that they are asked to memorize. The teachers' role is that of facilitator in the students' development. The aim is that assessment information becomes a scaffolding mechanism for instructional change that further facilitates student learning. The long-term goal is for teachers to become autonomous assessors of their students' understanding of science, which will be evidenced by their own professional development in terms of changing instructional and assessment practices.

The Challenge of Multilevel, Multisite Evaluation

In the investigation of teacher change in SEPUP's implementation of *Issues, Evidence and You*, it is clearly a case in which there are multiple levels of effects. SEPUP began as a multisite program with 15 Professional Development Centers (PDCs) across the United States[1]. These Centers were both internally and externally unique from each other. PDCs were internally different depending on the locus of leadership (school, district or university), and whether or not the teachers were from the same school and/or district. Externally PDCs were unique because of the varying state and local mandates placed on teachers within the centers (e.g., state testing and/or district curricular mandates). For the field test and evaluation, six of the original PDCs were designated Assessment Development Centers (ADCs) and were required to use the assessment system as they taught *IEY*. Seven sites continued as PDCs testing only *IEY* and two PDCs chose not to continue. Students were nested in SEPUP classrooms and teachers were nested in either ADCs or PDCs.

Further, the program affects both teachers and students because *IEY* represents a complete replacement of the typical middle school science course. The teachers are expected to teach using an issue-oriented science approach and the students are expected to do hands-on science in small, cooperative groups. The roles for both are significantly different than in the traditional middle school science course.

In a multisite evaluation, there is the potential for a program-by-site interaction (Sinacore & Turpin, 1991), so that simply pooling data to determine an overall program effect may lead to flawed interpretations (Keppel, 1991). Qualitative findings from site visits and teacher interviews indicated that there are Center differences on several factors, so the nested structure of the evaluation could not be ignored.

Cooley, Bond and Mao (1981) suggest that researchers faced with hierarchical data structures begin the multilevel analysis by understanding the relationship at either the higher (e.g., teacher to center) or lower (e.g., student to teacher) level first. They also emphasize the need to choose a causal model prior to choosing a method of analysis, pointing out that "the common practice of analyzing a multilevel set of predictors using a single multiple regression equation makes it extremely unlikely that one will find an effect for the higher level variables" (Ibid., p. 70). This is because the higher-level variable is constant within each lower level analysis, therefore within-center variation cannot be explained by center-level variables. However, the Hierarchical Linear Modeling (HLM) statistical approach can help overcome this difficulty (Bryk & Raudenbush, 1992; Seltzer, 1995) if the sample size at both the individual and group levels is sufficiently large.

HLM is a fairly recent innovation (Bryk & Raudenbush, 1993; Raudenbush & Bryk, 1986) which statistically confronts the problem of identifying the effects of educational programs or innovations from multilevel data. In the conclusion of his seminal article on multilevel data analysis, Burstein (1980, p. 223) noted it is

important to view the multilevel structure of data as a benefit, because the effects of educational processes need to be viewed from both the individual and the group perspective. Otherwise educational research and theories will not align with reality.

In the SEPUP case, HLM could not be used due to insufficient sample size. For this study, a cross-level exploratory approach was used to study teachers nested within Centers and to examine the impact of certain Center characteristics. Variations within Centers, such as quality of leadership, were hypothesized to affect teacher professional development in terms of changing assessment, collegial and instructional practices.

Seltzer (1992) discussed the value of linking quantitative and qualitative research traditions in multisite evaluations through the use of multilevel modeling. One benefit of this mixed-method approach is that the qualitative information collected from multiple sites can be used to help explain discrepant cases or outliers. The cases that generate the most interest are those that are either unusually successful in using an innovation or those that encountered unusual levels of failure or less than anticipated progress. A second benefit is the richness of detail provided by the qualitative information, which when triangulated with the quantitative evidence helps to explain the reasons for success or failure that otherwise might remain beyond the grasp of the evaluation researcher. In the SEPUP case, mixed methods were used to capitalize on these benefits.

Local Assessment Moderation

In SEPUP, local assessment moderation was used to provide high quality professional development opportunities for the teachers using the assessment system. Teachers engaged in local assessment moderation function as a community of judgment, and this task engagement serves as a medium for teacher change. Principles of high quality professional development, such as ongoing support rather than one-shot experiences (Fullan, 1991) and "support for informed dissent" (Little, 1993, p. 138), delineated by these and other researchers (e.g., Sparks & Loucks-Horsley, 1990) were adhered to in the design of local assessment moderation for SEPUP. Local assessment moderation is a critical element of the SEPUP Assessment System as it supports teacher professional growth in the politically charged arena of assessment reform.

An ongoing evaluation of the implementation of the consensus moderation process in Victoria was conducted from 1981, when it was first implemented, to 1984, with a follow-up study in 1989. Based on the four year evaluation in Victoria, Ingvarson (1990, p. 9) noted "the importance of regarding moderation as a complex innovation requiring a considerable period of time for [teachers'] learning and unlearning during its implementation." The Victoria study indicates that the consensus moderation process had "impressive side effects on the professional development and accountability of teachers" (p. 2). Many of the comments from the

SEPUP ADC teachers are consistent with the findings of the Victoria evaluation (Ingvarson, 1990). In the Victoria evaluation, it was found that involvement in the consensus moderation process:

(a) added significantly to teachers' skills for assessing student learning;
(b) enhanced teachers' ability to evaluate and improve their teaching;
(c) significantly increased teachers' access to useful ideas for teaching;
(d) enhanced the quality of learning of students;
(e) affected positively participants' teaching in non-project classes; and
(f) supported, rather than intimidated, beginning teachers.

Ingvarson (1990) also reported that the positive responses increased as teachers had more experience with moderation, which again reflects the need for time to become knowledgeable and skilled in using this process. This experience seems to hold true for at least one of the ADCs that had used the moderation process during both the pilot and field test years. This particular ADC's director reported that the teachers have changed. They have begun to appreciate the moderation process more. They used moderation to inform instruction and are now convinced of its instructional value. This director noted that in the first year some of the teachers were making up their own multiple choice tests, but that during the field test they "didn't use the old assessments anymore."

Methods

I refer to the methodology used to evaluate SEPUP as *embedded evaluation*. Embedded evaluation can be defined as a systemic effort to evaluate a teacher professional development program that is iterative, autonomous, and evidence-based. This evaluation methodology is *iterative* because the information collected at any one level serves to evaluate the next level(s). For example, the assessment of student learning contributes to an evaluation of teacher enhancement, and information about student achievement and teacher change can be used to evaluate the overall project. Embedded evaluation is *autonomous* because the key individuals at each level are increasingly responsible for the design, collection and evaluation of information as they develop professionally. This evaluation process provides *evidence* based on the project's goals as understood by the participants, or change agents, at each level.

Evaluation Research Questions

The research questions correspond to the multilevel nature of the SEPUP evaluation. The first two questions focus on the individual level, examining the effect that program participation has had on teachers. The last two questions focus on the group level, specifically looking at the teachers in the Assessment Development Centers (ADCs) who used the assessment components and

participated in local assessment moderation. These latter two questions examine the organizational factors (i.e., ADC features) that are associated with teacher change.

Individual level.

1.1 Were the changes for the SEPUP teachers greater than those for the non-SEPUP teachers in terms of assessment practices, instructional practices, and collegiality?

1.2 Were the changes for the SEPUP ADC teachers greater than those for the SEPUP PDC teachers in terms of assessment practices, instructional practices, and collegiality?

Group level.

2.1 What features of the ADCs are associated with differences in teacher change among Centers?

2.2 How do the ADCs differ in terms of teacher change?

Design of the Study

A mixed methods research design was used. Both quantitative and qualitative information was collected for the following three reasons: (a) to be able to triangulate quantitative and qualitative findings given the developmental nature of the SEPUP program; (b) to enable understanding of outliers or discrepant cases; and (c) to identify features of the Assessment Development Centers (ADCs) that had a direct effect on the use of the assessment system and consequently on teacher professional development.

The value of using qualitative procedures to complement quantitative methods in this evaluation hinges on the fact that the SEPUP Assessment System was under development and being field tested. Qualitative research can provide a rich source of information in which to look for alternative explanations, examine patterns, verify findings, and as needed, formulate new research questions (Miles & Huberman, 1994). The quantitative information summarizes the results numerically, allowing for judgments as to the degree of program impact. The combination of these two research traditions leads to a powerful means to examine program impact, question results and identify potential factors or variables to be studied further.

Subjects.

To evaluate SEPUP during the field test, three treatment groups were defined: SEPUP teachers in Assessment Development Centers (ADCs); SEPUP teachers in Professional Development Centers (PDCs); and non-SEPUP, comparison teachers, at

least one from each of the Centers. All of these teachers taught middle school science. The SEPUP teachers used the *IEY* curriculum and the non-SEPUP teachers used their traditional science curriculum. The SEPUP ADC teachers received the most comprehensive treatment, both the curriculum and the assessment tools designed for SEPUP.

The sample consists of all teachers involved in the 1994-95 field test of SEPUP's *Issues, Evidence and You*. A portion of these teachers had also participated during the pilot year, but about a third of the SEPUP teachers were new to the program in the field test year. There were 26 SEPUP teachers and seven comparison teachers in the ADCs. There were 25 SEPUP teachers and five comparison teachers in the PDCs.

Only the four ADCs that complied with all aspects of the SEPUP Assessment System field test were included in the analyses presented here. Two ADCs were dropped because the teachers did not fulfill their obligations with regard to local assessment moderation and the return of evaluation information.

The participating teachers were from various schools and districts within the Centers. Centers were organized a bit differently in each location. Some Centers were located in a single, large school district and a district representative functioned as the ADC director. Other Centers were organized less centrally, for example, one ADC had five SEPUP teachers from five different districts, and one of these teachers also served as the ADC director. In some cases, the Center director was not from the local school or district, but rather a university person involved in science education.

Description of the instrument.

Instruments were developed to meet the particular needs of this evaluation research following an exhaustive review of the literature which did not uncover appropriate instruments. The teacher-level instrument discussed in this paper is the *SEPUP Inventory of Teachers' Assessment, Collegial, and Instructional Practices* (SITACIP). The initial SITACIP was pilot tested in 1993-94, then a modified version was used in 1994-95. For the validation study refer to Roberts (1996).

The SITACIP survey is a 77-item self-report of teachers' assessment, instruction and collegial practices as well as their perceptions about the usefulness of various assessment strategies for assessing learning, guiding instruction and grading. The subsections of the SITACIP fall into three basic question types: (a) frequency of use of various instructional and assessment strategies; (b) Likert-type scales for both collegiality and reasons for assessment strategy choices; and (c) attitudes about the usefulness of different assessment strategies for assessing learning, guiding instruction, and grading.

Table 1 presents the four SITACIP scales that are described in this paper. The scales measure teachers' attitudes, perceptions and practices related to assessment, instruction, and collegiality.

INSERT TABLE 1 ABOUT HERE

Data Analysis

The quantitative data were analyzed in two veins: first for descriptive purposes and second to make statistical inferences. The statistical techniques used for descriptive purposes were frequency counts and central tendency statistics (e.g., the mean). The statistical techniques used for inferential purposes include: (a) t-tests to examine pre to post mean differences within-Centers; and (b) Analysis of Covariance (ANCOVA) using the pretest scale score as the covariate to compare treatment groups (e.g., ADC vs. PDC teachers). The independent variables came from a teacher background survey and project records. The dependent variables were the post-SITACIP scale scores for assessment, collegial and instructional practices.

Table 2 presents the type of statistical analyses that were conducted by research question, the comparison or purpose for the analysis, and the instrument or data source. Explanations of the various analyses follow the table.

INSERT TABLE 2 ABOUT HERE

Before addressing the first research questions, t-tests were used to examine pre to post mean differences on the SITACIP scales for the three groups (ADC, PDC and comparison teachers). These differences were used to identify the size and direction of change. The acceptable minimum level of significance was set at $p < .05$. The analyses were only done for matched cases (i.e., the pre and post score was available or the teacher was dropped from the analysis).

ANCOVA was used because of limitations imposed by the purposive sampling. There was no possibility under the circumstances to use a block design, therefore a procedure that statistically "blocks" the participants was the most appropriate procedure to use (Keppel, 1991). In this case, the pre-scale scores on the SITACIP were used as the covariates when comparing two groups of teachers. Before making group comparisons, the assumption of homogeneity of slopes or group regression coefficients (Keppel, 1991, p. 316-321) was tested. The purpose of this test is to determine whether there is a significant interaction between the covariate and the groups (e.g., ADC and PDC teachers). For the assumption of homogeneity to hold, the slope of the regression line for the dependent variable

onto the covariate is supposed to be the same for both groups. In the case of significant differences or heterogeneity, the effect of the independent variable needs to be interpreted with caution.

Using ANCOVA, the first comparison made was the difference on the post-scale scores between all SEPUP teachers (i.e., both ADC and PDC) and the comparison teachers. The next comparison made was the difference between the two types of SEPUP teachers (i.e., ADC vs. PDC) on the post-scale scores. The acceptable minimum level of significance was set at $p < .05$. The adjusted multiple R squared is presented with the ANCOVA results as the effect size which represents the proportion of variance explained by the group variable.

Teacher-level Results: Comparing the Impact of Different Treatment Levels

First, a summary of the within-group t-tests for each of the three groups (ADC, PDC and comparison teachers) on the SITACIP scales is presented. Next, analysis of covariance (ANCOVA) results for the SITACIP scales are presented in summary form. The ANCOVA results address the first two research questions by examining the differences between SEPUP and comparison teachers as well as ADC and PDC teachers.

Evaluating Teacher Change: Summary of the Quantitative Evidence

Measuring and evaluating teacher change is not an easy task. Change is a process that takes time and may not be captured in the short time frame of a funded project. This quantitative evidence does not sufficiently capture the nuances of teacher change. Nevertheless, quantitative evidence provides a starting point from which to explore teacher change further.

Analysis of Within-Group Differences.

The pre to post mean differences in Table 3 (under t-tests) begin to paint a picture of SEPUP's effect on teachers' professional development in the areas of assessment, collegial or instructional practices. Although largely non-significant, the directionality of change for the two groups is of great interest to those concerned with teacher change and assessment reform. The ADC teachers had a slight decrease on the Assessing Learning scale whereas the PDC teachers had significant increases on this same scale. However, PDC teachers had a decrease on the Assessment Strategy Use scale by the end of the field test. Collegiality increased more for the ADC than the PDC teachers, but neither gain was significant. The ADC and PDC teachers had no significant differences on the Instructional Activity Use scale -- this was to be expected given that these two groups were using the same curriculum and received the same treatment in terms of instructional strategies and models for delivering instruction.

INSERT TABLE 3 ABOUT HERE

On the scales for which matching data exists for comparison teachers, this group did not change significantly. Since the subsequent analyses all indicate that the non-SEPUP teachers had no significant changes on any of the SITACIP scales, this group is not considered further in the presentation of results.

If the evaluation of SEPUP were to stop at this point, the program might appear to have been marginal in facilitating teacher enhancement. Specifically, the SEPUP Assessment System might be considered ineffective in changing ADC teacher's collegial, instructional and assessment practices. There was a positive, albeit non-significant (due most likely to the small sample size and concomitant loss of statistical power), change on the Collegiality scale for the ADC teachers overall. Further, ADC teachers increased their use of alternative assessment strategies while PDC teachers decreased their use of such strategies.

The fact that the SEPUP Centers are located in several states and numerous school districts across the U.S. raises the issue of between-Center differences. Do these aggregate results mask or significantly dilute the program's effects in different settings? Are there exogenous factors that mediate the program's effects within different Centers? These questions can only be addressed using multilevel evaluation.

Analysis of Between-Group Differences.

Prior to using ANCOVA, the assumption of homogeneity of slopes was tested. Only the Assessing Learning scale had a significant interaction effect between the treatment groups (ADC and PDC teachers) and the covariate. The results for this comparison were interpreted with caution as advised by Keppel (1991). To inspect for interaction, the covariate means and the dependent variable (post score) means were plotted for each group for this scale. The plot in Figure 1 clearly shows the significant interaction effect. Note that the regression lines crisscross because the ADC teachers on average had a lower post score than pre score on the Assessing Learning scale while the PDC teachers had a larger post mean than pre.

INSERT Figure 1 ABOUT HERE

The right-hand column on Table 3 summarizes the ANCOVA results comparing the ADC and PDC teachers. In Table 3, the adjusted multiple R squared is presented as the effect size which represents the "proportion of adjusted variability" explained by the group (Keppel, 1991, p. 322). Using Cohen's (as cited in Keppel, 1991, p. 66) description of effect sizes for the behavioral and social sciences,

the adjusted multiple R squared statistics presented in Table 3 can be interpreted as follows: a small effect size is about .01 omega squared ($\omega^2_A = \sigma^2_A / \sigma^2_A + \sigma^2_{S/A}$), a medium is approximately .06, and a large is about .15. Keppel notes that R squared is always larger than omega squared, but Cohen's estimates of effect size can still be used to interpret results. Reporting the adjusted multiple R squared from the ANCOVA in addition to the F statistic is critical. The effect size measure is not unduly influenced by the sample size. Consequently, under low power created by small samples, as is the case here, the effect size can be large and go undetected by the F statistic. The F statistic is insensitive in the case of small samples and low power. For example, the Assessing Learning scale was not statistically significant when comparing ADC and PDC teachers. However, the adjusted multiple R squared for Assessing Learning is .18, indicating that this is a large effect size. Hence, the F statistic failed to detect this large difference due to the low power created by the small, purposive sample. However, the effect size was much larger ($R^2 = .31$) for the Assessment Strategy Use scale, and this did result in a significant ANCOVA when comparing ADC and PDC teachers.

On the Assessment Strategy Use scale, ADC teachers changed significantly more than PDC teachers. Note in Table 3 that the ADC teachers had a positive change pre to post on this scale while the PDC teachers had a decrease by the end of the school year. On the Assessing Learning scale, the opposite direction of change occurred (i.e., the ADCs went down and the PDCs went up significantly); this difference between ADC and PDC teachers was almost significant ($p = .08$).

Center-level Results: Exploring Teacher Change within ADCs

This section presents the last steps in one solution to the methodological challenge of multilevel, multisite evaluation when the sample size at the group level is fairly small. Contextual factors that affected the implementation of local assessment moderation were identified using qualitative methods. Consequently, I was able to order the ADCs by level of success with local assessment moderation and then look at the quantitative findings across Centers through a more focused lens. Using a cross-level exploratory approach, I was able to evaluate SEPUP and avoid pitfalls, such as aggregation bias.

Evaluating Multiple Sites: Summary of the Qualitative Evidence

There are many benefits to be had by teachers engaging in local assessment moderation, but there are barriers and limitations that were also discovered as the implementation of the SEPUP Assessment System was evaluated. Both the benefits and the potential barriers to successful implementation are discussed below.

Contextual Factors that Affected Implementation.

The level of success with implementation of the local assessment moderation was determined to a large extent by the organizational context; in other words, the Center mattered. Features of the ADCs that mattered include: strength and quality of leadership; institutional support; and teacher proximity and collaboration (see Figure 2 for a summary of findings by Center).

INSERT FIGURE 2 ABOUT HERE

Leadership for Change. The ADC teachers in general came from different schools, so their experiences with "norms of collegiality" (Little, 1982; McLaughlin, 1993) varied. Some teachers felt isolated in their schools while others participated in science department or grade level teams. The key to bringing the group together cohesively was the leadership provided by the ADC director or by strong teachers within the group who had prior SEPUP experience. The importance of leadership runs the gamut of ensuring that the group develops a rapport to preparing teachers to use local assessment moderation to facilitating moderation sessions.

Who the leader is may not be all that important, but according to the ADC teachers, there needs to be someone in a leadership position to move the group toward consensus, to intervene in personal conflicts, to diffuse philosophical differences that digress from the work at hand, and overall to keep the group on task. Time, as always, is a factor for teachers, so maintaining task orientation is important.

Institutional support. As Little (1993) and others have suggested, a quality professional development program needs to balance support for institutional initiatives with support for those initiated by teachers. For some of the ADCs, the purpose of changing teachers' assessment practices was consistent with state initiatives or local school or district-level staff development goals. These goals were balanced with the needs of the participating teachers, who for the most part were very successful science teachers, but were interested in learning about new assessment strategies. In terms of school-based administration, teachers need support from their principals, and in the case of SEPUP this was generally not a problem. In fact, many of the SEPUP teachers have a long history of involvement in innovations and have been supported by their principals.

Strong leadership was also important in securing district-level support. The most critical support factor related to the district was gaining access to the ADC teachers for whole-day or half-day meetings rather than after-school sessions, including release time[2] and substitutes. Having sufficient time as a group was important for three fundamental reasons; time is needed: (1) to build a group rapport or a SEPUP norm of collegiality; (2) to build teacher understanding of the

SEPUP assessment system; and (3) for teachers to function as reflective practitioners. Having district support has broader implications as well, but with time being an oft-noted issue for teachers, it is critical to not lose sight of the opportunity cost of teachers' time. In two of the ADCs, the ADC director was a representative from the district.

Teacher Proximity and Collaboration. Having teachers from more than one district can work. In ADC 2 (refer to Figure 2), two teachers from a smaller district, an hour's drive north of the larger participating district, communicated with each other between moderation meetings, while the teachers in the larger district tended to call the ADC Director (who worked at the District office) or one of the experienced SEPUP teachers. Given the long commute to attend meetings for three of the teachers, this ADC met for whole days. These two districts had an outstanding history of collaboration. However, I also found that too many districts can be detrimental and minimize the leadership of the ADC director (e.g., ADC 1).

Teacher collaboration is related to proximity in that teachers who are closer tend to be able to spend more time communicating with one another. One of the ADC 2 teachers noted that she would have liked to confer with one of the teachers from the other district, but that it was a long distance call, so she tended not to contact her between meetings. Teachers most often called one another to ask a question about an assessment task or to clarify something on a scoring guide. On occasion, teachers would actually meet or pursue a mutual project together.

Based on site visit observations, the quality of moderation varied, but so did the amount of experience with using local assessment moderation and the amount of time devoted to staff development. As noted above, the full days that some ADCs used were critical not only to forming a cohesive group, but to learning about the moderation process and learning from each other how the assessments were working in their classrooms. In this way, the moderation meetings became the ongoing support that teachers needed as they learned about and implemented the SEPUP course and assessment system.

Evaluating Multiple Sites: Summary of the Cross-Level Exploratory Analyses

Bryk and Raudenbush (1993) recommend pursuing cross-level exploratory analyses when examining multilevel data. In this way, patterns may emerge that guide further analyses. The approach outlined below provides a view of teacher change across ADCs, thus opening up a discussion about within-Center differences. Further, this approach is best used when sufficient qualitative information is available in addition to quantitative information, therefore a mixed-method approach is recommended, especially when the number of groups is fairly small.

Figure 3 presents the mean difference change plot for the Assessment Strategy Use scale. The mean difference in logits for each ADC is noted in parentheses beneath the plot and the mean is represented by a dark circle. Lines connect the

mean differences, indicating the pattern of differences across ADCs and in relationship to the overall mean difference for the four ADCs. The ADCs are arranged from least (ADC 1) to most (ADC 4) successful with local assessment moderation based on the qualitative evidence.

INSERT FIGURE 3 ABOUT HERE

Only one ADC had a decrease on the Assessment Strategy Use scale (ADC 4). One teacher in ADC 4 had a very high pre-score which dropped more than one logit to just above the post mean for all four ADCs. The other three teachers began with scores closer to the pre-scale mean for all ADCs. Two of these teachers had only slight decreases, but one of them decreased by nearly one logit.

All the other ADCs had increases on the Assessment Strategy Use scale. Two beginning teachers (i.e., new to teaching) in ADC 3 had very large gains on the Assessment Strategy Use scale, indicating that they reported using alternative assessment strategies much more frequently than most other ADC teachers by the end of the field test. The overall post mean indicates that teachers most likely use alternative assessments about once a week, which sounds reasonable given the nature of the assessment activities in *IEY* (e.g., lab reports). Given the probationary status of new teachers, these responses could reflect overly ambitious assessment use.

Figure 4 presents the plot of standard errors for the post-scale means for the four ADCs on the Assessment Strategy Use scale. The standard errors provide a more fine-grained interpretation of the mean differences. The post scale means are represented by ellipses. Lines extend from the means to plus and minus one standard error represented by a thick horizontal bar. All the post standard errors are fairly small, indicating that the sample mean is a good approximation of the expected value. ADC 3 appears to be an outlier in the plot in Figure 4. The qualitative evidence regarding the two beginning teachers assists in explaining this discrepant case.

These plots of the mean differences and the standard errors provide one solution to the interpretation of multilevel, multisite evaluation data. The previous example indicates how the combination of quantitative and qualitative evidence can provide substantively interpretable information.

Discussion

One of the prominent themes that emerges in the evaluation results is the difference between the rhetoric and the reality of assessment reform as it plays out in SEPUP teachers' minds and classrooms. The SEPUP ADC teachers were intensively engaged in testing the components of the SEPUP Assessment System while the SEPUP PDC teachers were only testing the curriculum. The ADC teachers, while embracing the rhetoric of alternative assessment reform in the beginning, changed their perceptions of the usefulness of various alternative assessment strategies over time. Faced with the reality of implementing alternative assessments, ADC teachers modified their perceptions of the usefulness of such strategies for assessing learning, guiding instruction and grading. ADC teachers continued to feel that alternative assessments were valuable, but were more reserved in their perceptions. ADC teachers did increase their use of open-ended questions while significantly decreasing the use of closed-ended questions. ADC teachers' assessment practices did change, but their attitudes seemed to be in a state of dissonance as they tried to reconcile -- through practice -- the rhetoric and the reality of assessment reform.

Meanwhile, the PDC teachers still embraced the rhetoric of reform by the end of the field test, and yet their assessment practices had retreated to a traditional stance. Even though the *IEY* course has embedded assessment tasks, the results indicate that the PDC teachers resorted to very traditional modes of assessment. In particular, they increased their use of closed-ended questions by the end of the field test year. PDC teachers did have access to the scoring guides for the SEPUP variables, but it seems that access to scoring guides did not lead to use of the embedded assessment tasks. This traditional approach to assessment may explain the variation in ADC and PDC student achievement results reported elsewhere (Wilson & Draney, 1996).

Student achievement is another indicator for evaluating the success of SEPUP's *IEY*. Although, student achievement was not included in the multilevel analysis reported here, student achievement on the SEPUP variables is worth noting from both a policy and an evaluation perspective. Students in ADC teachers' classes exhibited significant growth over the field test year on the SEPUP assessment tasks, pre/post tests and link tests. However, the PDC and comparison students had comparable patterns of non-significant growth. In other words, the slope of the growth line was much steeper for the ADC students while virtually flat for the PDC and comparison students.

For the dissemination of the *IEY* course with an embedded assessment system, these results may have serious consequences. In light of the rhetoric for assessment reform, the embedded assessment may be the selling point of the *IEY* course, but without inservice and ongoing support for teacher change, the very heart of the course that is being valued might not even be used.

In conclusion, I set out to address the common methodological challenge presented by the multilevel structure of complex innovations. I used cross-level exploratory analyses as a means to interpret between-Center differences. Further, I paid close attention to effect sizes rather than F statistics because of the small sample size which compromised the statistical power of the analyses. Without the qualitative information, I would have been at a loss to explain the mixed impact of SEPUP at the Center level. The richness of detail provided by site visits and through personal interviews and focus groups allowed me to order the ADCs and look at between-Center differences through a more focused lens. I feel strongly that the value of using mixed-methods has been underscored in this example of an evaluation of a complex innovation.

These findings illuminate the difficulties in multilevel evaluation and yet present positive steps one can take to meet this challenge. As Burstein (1980) noted, the reality of educational research is that it is multilevel, and consequently to make meaning of what we explore we must move forward in examining these real-world structures to the best of our ability. At this point, the statistical technology cannot accommodate problems such as small sample sizes in multilevel, multisite evaluations. However, I see the approach outlined here as being helpful for at least formative evaluation (which is how it was used) and perhaps for summative evaluation.

Notes

1. States with PDCs (states that had ADCs during the field test are in bold): **AK**, **CA**, **CO**, **KY**, **LA**, MI, **OK**, NY, NC, and PA.

2. The assessment blueprint is an integral part of the SEPUP Assessment System. The blueprint details the placement of assessment tasks throughout the course as well as identifying the course variables that can be assessed.

3. The ADCs received a budget for participation, but some chose to pay teachers' stipends and meet after school, while others paid for substitute costs with the district providing teacher release time.

Reference List

American Association for the Advancement of Science (AAAS), Project 2061. (1993). Benchmarks for science literacy. New York: Oxford University Press.

Blumenfeld, P.C., Krajcik, J.S., Marx, R.W., & Soloway, E. (1994). Lessons learned: How collaboration helped middle grade science teachers learn project-based instruction. The Elementary School Journal, 94 (5), pp. 539-551.

Bryk, A.S., & Raudenbush, S.W. (1992). Hierarchical linear models: Applications and data analysis methods. Newbury Park, CA: Sage.

Bullough, Jr., R.V. (1988). Excellence and testing. Ch. 4 in The Forgotten Dream of American Public Education. Ames: Iowa State University Press.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. In Review of Research in Education, 8, 158-233.

Cooley, W.W., Bond, L., & Mao, B. (1981). Analyzing multilevel data. In R.A. Berk (Ed.) Educational Evaluation Methodology: The State of the Art (pp. 64-83). Baltimore: The Johns Hopkins University Press.

Fullan, M. (1991). The new meaning of educational change (2nd edition). New York: Teachers College Press.

Harmon, M. (1995). The changing role of assessment in evaluating science education reform. In R.G. O'Sullivan (Ed.) Emerging roles of evaluation in science education reform (pp. 31-52). New Directions for Program Evaluation, Number 65. San Francisco, CA: Jossey-Bass Publishers.

Ingvarson, L. (1990). Enhancing professional skill and accountability in the assessment of student learning. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA. (ERIC Document Reproduction Service No. 327 558)

Jett, D.L., & Schafer, W.D. (1992). Classroom teachers move to center stage in the assessment area--ready or not! Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. 346 144)

Keppel, G. (1991). Design and analysis: A researcher's handbook. Englewood Cliffs, NJ: Prentice Hall.

Krajcik, J.S., Blumenfeld, P.C., Marx, R.W., & Soloway, E. (1994). A collaborative model for helping middle grade science teachers learn project-based instruction. The Elementary School Journal, 94 (5), 483-497.

Levine, H. (1994, May). Interactive mathematics project: Networking and implementation. In St. John, M., Levine, H., Håkansson, S., Lopez-Freeman, M., Roberts, L., & Shattuck, J. A Study of the California Dwight D. Eisenhower Mathematics and Science Education State Grant Program. Inverness, CA: Inverness Research Associates.

Linn, R.L. (1987). Accountability: The comparison of educational systems and the quality of test results. Educational Policy, 1 (2), pp. 181-198.

Linn, R.L. (1994). Performance assessment: Policy promises and technical measurement standards. Educational Researcher, 23 (9), 4-14.

Little, J.W. (1982). Norms of collegiality and experimentation: Workplace conditions of school success. American Educational Research Journal, 19 (3), 325-340.

Little, J.W. (1993). Teachers' professional development in a climate of educational reform. Educational Evaluation and Policy Analysis, 15 (2), 129-151.

Lopez, R.E., & Tuomi, J. (1995). Student-centered inquiry. Educational Leadership, 52 (8), pp. 78-79.

McLaughlin, M.W. (1993). What matters most in teachers' workplace context? In J.W. Little & M.W. McLaughlin (Eds.), Teachers' work: Individuals, colleagues, and contexts (pp. 79-103). New York: Teachers College Press.

Miles, M.B., & Huberman, A.M. (1994). An expanded sourcebook: Qualitative data analysis (2nd ed.). Thousand Oaks, CA: Sage Publications.

National Council on Measurement in Education (NCME). (1994). Conference Highlights Keys to Better Assessment. NCME Quarterly Newsletter, 3 (1). Washington, D.C.: National Council on Measurement in Education.

National Council of Teachers of Mathematics (NCTM). (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: NCTM.

Newell, S.T. (1992). Science teachers' perspectives on alternate assessment. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1992.

Office of Technology Assessment. (1992). Testing in American Schools: Asking the Right Questions. Washington, D.C.: Congress of the United States.

O'Sullivan, R.G. (Ed.) (1995). Emerging roles of evaluation in science education reform. New Directions for Program Evaluation, Number 65. San Francisco, CA: Jossey-Bass Publishers.

Pennell, J.R., & Firestone, W.A. (1996). Changing classroom practices through teacher networks: Matching program features with teacher characteristics. Paper presented at the annual meeting of the American Educational Research Association, New York, NY, April.

Raudenbush, S., & Bryk, A.S. (1986). A hierarchical model for studying school effects. Sociology of Education, 59 (January), 1-17.

Richardson, V. (1990). Significant and worthwhile change in teaching practice. Educational Researcher, 19 (7), pp. 10-18.

Roberts, L. (1995a). Assessment as a vehicle for teacher professional development: Moving teachers further along the continuum. In St. John, M., Levine, H., Håkansson, S., Lopez-Freeman, M., Roberts, L., Shattuck, J., & Von Blum, R., A Study of the California Dwight D. Eisenhower Mathematics and Science Education State Grant Program: Lessons Learned about Designing Professional Development Projects, (Essay V). Inverness, CA Inverness Research Associates.

Roberts, L.L.C. (1996). Methods of evaluation for a complex treatment and its effects on teacher professional development: A case study of the Science Education for Public Understanding Program. Unpublished dissertation, University of California, Berkeley.

Roberts, L. (under review). Using maps to produce meaningful evaluation measures: Evaluating changes in middle school science teachers' assessment perceptions and practice. Submitted to the editors for Objective Measurement: Theory into Practice (Vol. 5).

Seltzer, M.H. (1995). Furthering our understanding of the effects of educational programs via a slopes-as-outcomes framework. Educational Evaluation and Policy Analysis, 17 (3), 295-304.

Seltzer, M.H. (1992). Linking the quantitative and qualitative traditions in multisite evaluations: A multilevel modeling approach. Paper presented at the annual meeting of the American Evaluation Association, November 1992, Seattle, WA.

Shepard, L.A. (1995). Using assessment to improve learning. Educational Leadership, 52 (5), 38-43.

Shepard, L.A. (1989). Why we need better assessments. Educational Leadership, 46 (7), pp. 4-9.

Sinacore, J.M., & Turpin, R.S. (1991). Multiple sites in evaluation research: A survey of organizational and methodological issues. In Multisite Evaluations (pp. 5-18), New Directions for Program Evaluation (50). San Francisco: Jossey-Bass, Inc. Publishers.

Sloane, K., Wilson, M., & Samson, S. (1996). Designing an embedded assessment system: From principles to practice. Paper presented at the annual meeting of the American Educational Research Association. New York, April.

Sparks, D., & Loucks-Horsley, S. (1990). Models of staff development. In W.R. Houston (ed.), Handbook of research on teacher education. New York: Macmillan Publishing Co.

Spring, J. (1988). Conflict of interests: The politics of American education. New York: Longman.

Tittle, C.K., Hecht, D., & Moore, P. (1993). Assessment theory and research for classrooms: From taxonomies to constructing meaning in context. Educational Measurement: Issues and Practice, 12 (4), pp. 13-19.

Tobin, K., & Gallagher, J.J. (1987). The role of target students in the science classroom. . Journal of Research in Science Teaching, 24 (1), pp. 61-75.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. Phi Delta Kappan, pp. 703-713.

Wilson, M. (1994). Community of judgment: A teacher-centered approach to educational accountability. In Office of Technology Assessment (Ed.), Issues in Educational Accountability. Washington, D.C.: Office of Technology Assessment, United States Congress.

Wilson, M., & Adams, R.A. (1996). Evaluating progress with alternative assessments: A model for Chapter 1. In M.B. Kane (Ed.), Implementing performance assessment: Promise, problems and challenges. Hillsdale, NJ: Erlbaum.

Wilson, M., & Draney, K. (1997). Developing maps for student progress in the SEPUP assessment system. Paper presented at the Middle School meeting of American Association for the Advancement of Science (AAAS). Seattle, WA, February.

List of Tables and Figures

Tables

Table 1. Description of SITACIP Scales

Table 2. Types of Statistical Analyses Used by Research Question

Table 3. Summary of t-tests and ANCOVA Results for selected SITACIP scales

Figures

Figure 1. Plot of the covariate and dependent variable means for the Assessing Learning scale

Figure 2. Level of success with local assessment moderation by features of the Assessment Development Centers

Figure 3. Change plot for the mean differences on the Assessment Strategy Use scale for the ADCs

Figure 4. Standard error plot for the post means on the Assessment Strategy Use scale by ADCs

Table 1

Scales	Type of Response Categories	Range of Responses
Assessing Learning - Perception of usefulness of different assessment strategies for developing an understanding of what students know.	Likert-type attitude scale (useless, somewhat useful, quite useful, very useful)	1 - 4
Assessment Strategy Use - Use of different assessment or evaluation strategies (e.g., lab reports; open-ended questions).	Frequency teacher uses each assessment or evaluation strategy (never to more than once per week)	1 - 5
Collegiality - Time spent discussing different topics with other teachers (e.g., assessment/ evaluation strategies).	Frequency teacher discusses topics with other teachers (often, i.e., never to more than once per month)	1 - 4
Instructional Activity Use - Use of different instructional strategies to teach middle school science (e.g., "You ask students questions about the material they've read or heard about.")	Frequency teacher uses each instructional strategy listed (never to every day)	1 - 5

Table 2

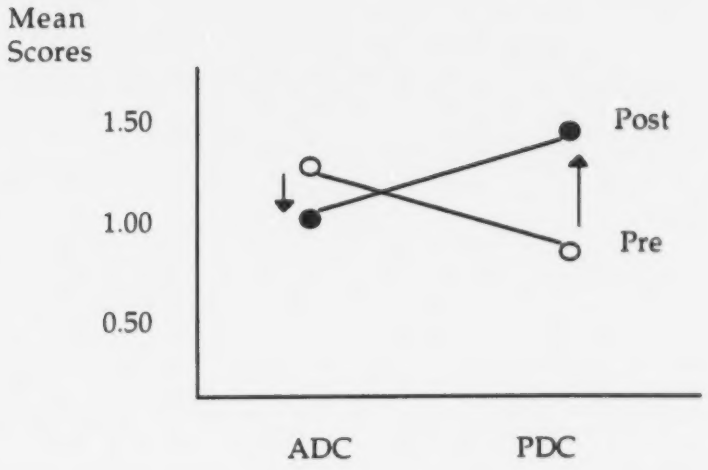
Instrument/ Source of Data	Research Question	Comparison/ Purpose	Statistical Analysis
SITACIP	1.1	SEPUP vs. Comparison	ANCOVA with pre-scale score as covariate
SITACIP	1.2	SEPUP ADC vs. SEPUP PDC	ANCOVA with pre-scale score as covariate
Focus Groups and Interviews	2.1	Identify features of ADCs associated with teacher change	No statistical analysis Case-ordered matrix
SITACIP	2.2	Identify how ADCs differ	Cross-level exploratory plots of pre/post mean scores

Table 3

Scale	t-test ADC			t-test PDC			ANCOVA ADC vs. PDC		
	n	Post[a]	M Diff.[b]	n	Post	M Diff.	df[c]	F[d]	Adj. R^2
Assessing Learning	10	1.00	-.22	12	1.40	.47*	19	3.34	.18
Assessment Strategy Use	19	.68	.20	11	.04	-.34	27	8.30**	.31
Collegiality	18	1.87	.33	11	1.62	.21	26	.11	.00
Instructional Activity Use	10	1.02	.58	13	1.24	.53	20	.18	.01

[a] Post: Post-test scale score in logits.
[b] M Diff.: Mean difference pre to post in logits.
[c] Degrees of Freedom associated with the S within group error.
[d] F ratio.
* $p < .05$.
** $p < .01$.

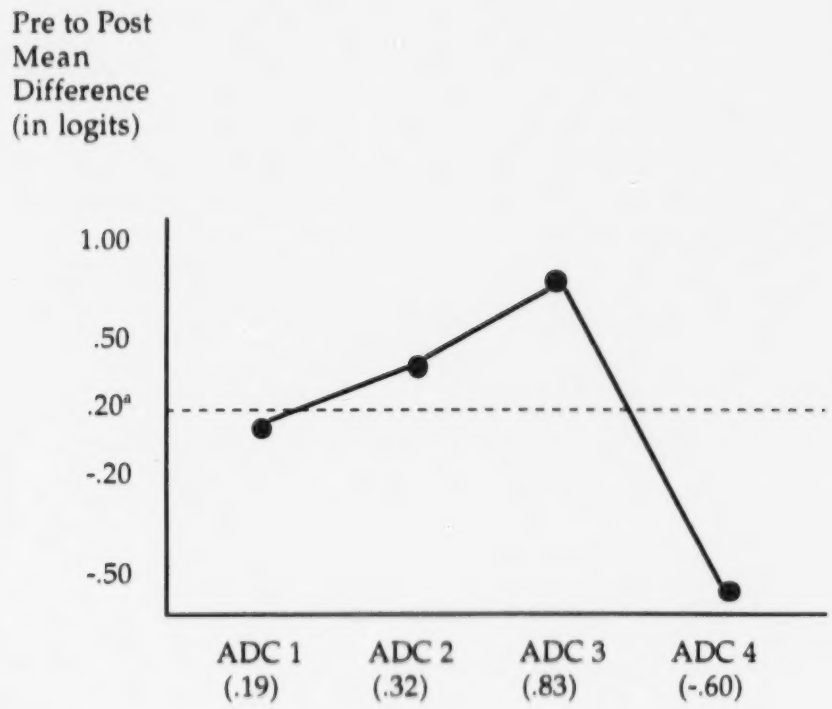
Figure 1



Level of Success with Implementation of Local Assessment Moderation	Leadership	Institutional Support	Teacher Proximity and Collaboration
Very High (ADC 4)	Strong leadership of ADC director. Former SEPUP teacher functioned as moderation facilitator.	Single district with district administrator as ADC director. Principals supported teacher release time. Efforts parallel State mandates for change.	All four teachers in the same district, but at different schools. Teachers work on other projects together. Teachers communicate via electronic mail or telephone.
High (ADC 3)	Teacher leader in place of ADC director who took on other duties.	Two schools in a single district. Principals supportive of staff development efforts.	Teachers paired at the two schools. Beginning teachers were supported by experienced teachers.
Moderate (ADC 2)	Two districts involved. Two teachers co-facilitated moderation meetings. The ADC director missed a few meetings due to other district obligations.	Districts emphasizing assessment as staff development focus. Have several federally-funded science education reform projects in the State.	Teachers who co-facilitated were accepted in this shared role by the other teachers. Teachers called others in the same district between meetings (otherwise long distance).
Limited (ADC 1)	Five teachers from five different districts. Limited by the fact that the ADC director was also a classroom teacher.	Varying district mandates diluted level of success. Principals' support varied, and there was no apparent district support. State drafting standards for science (K-12).	Limited somewhat by physical distance. Limited collaboration; called someone if a problem arose. Voice mail established by ADC director to provide weekly updates, but was not used by all teachers.

Figure 2

Figure 3



[a] Overall mean difference for all four ADCs. Mean Difference: ●

Figure 4

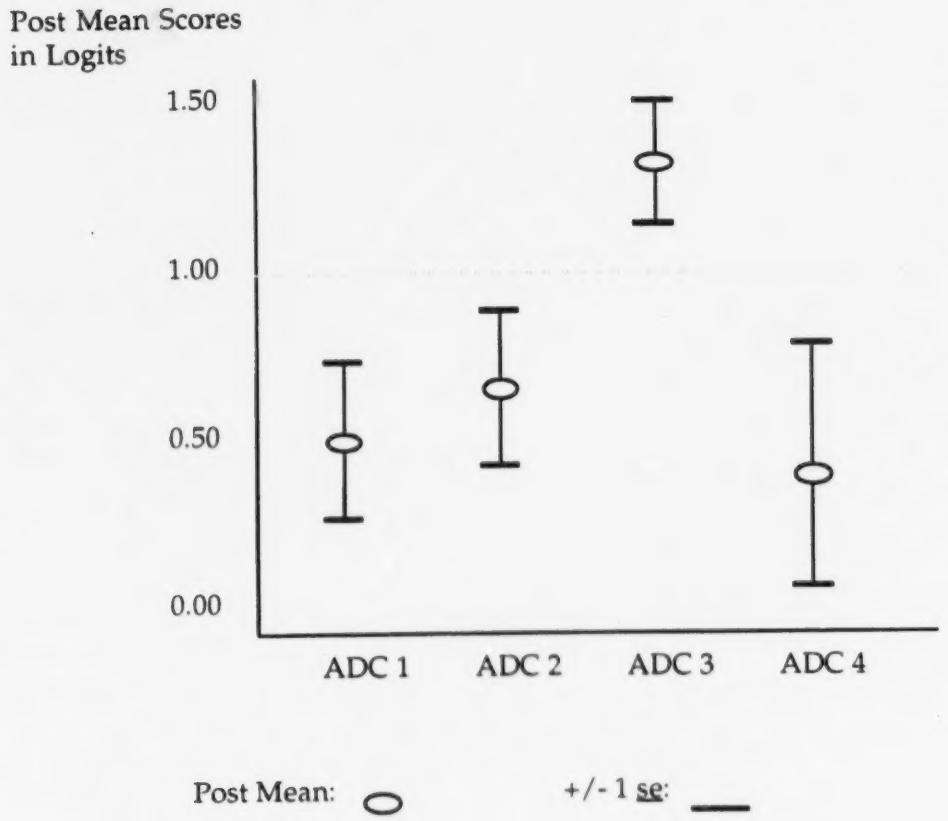
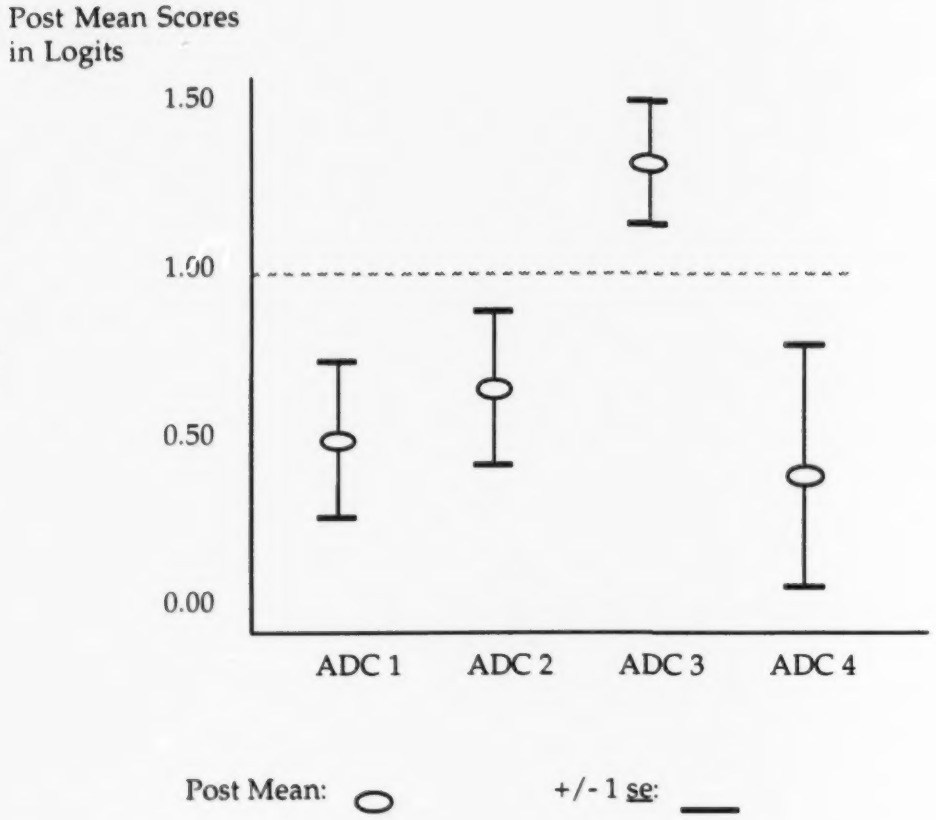

Post Mean Scores
in Logits
1.50
1.00
0.50
0.00
ADC 1
ADC 2
ADC 3
ADC 4
Post Mean:
+/- 1 se:

Figure 4


Post Mean Scores
in Logits
1.50
1.00
0.50
0.00
ADC 1
ADC 2
ADC 3
ADC 4
Post Mean:
+/- 1 se: